

WHAT HAPPENS WHEN YOUR SURVEYS NEED A RADICAL OVERHAUL

Carolyn F. Shettle, National Science Foundation

KEY WORDS: Surveys, research, quality

STPDS OF THE 1980s

INTRODUCTION

The primary goal of the Scientific and Technical Personnel Data System (STPDS) is to provide descriptive information about the number of U.S. Scientists and Engineers (S&Es), and their demographic, educational, and employment characteristics. A secondary goal is to provide researchers with data they can use to explore analytical goals, such as obtaining an understanding of labor market dynamics related to science and engineering.

From a technical point of view the largest problem facing the STPDS is that scientists and engineers constitute a fairly small part of the total U.S. population and there is not a comprehensive roster from which we can periodically select a sample of the population of interest. We, therefore, conduct a system of surveys, each of which is designed to provide information about a segment of the larger population of interest.

From an organizational point of view perhaps the greatest challenge facing the STPDS is that the Division of Science Resources Studies at NSF is a fairly small statistical agency, which does not have the resources to perform major surveys in-house. The actual surveys are, therefore, conducted under contract to NSF by three different contractors -- the Census Bureau, the National Research Council of the National Academy of Sciences, and a private contractor (most recently Temple University). In addition, we have a fourth contractor, Mathematica Policy Research (MPR), providing technical assistance for the surveys. Coordination among NSF staff, the survey contractors, MPR, and additional external experts who serve on our Advisory Board and/or provide assistance on specific tasks on an ad hoc basis is, therefore, an important challenge.

During the 1980s the STPDS had four major components. Three of the components were surveys of segments of the scientific and engineering population and the fourth component was a tabulation model which was used to estimate the total population from the parts.

The National Survey of Natural and Social Scientists and Engineers was the largest of the three surveys. The sample frame for this biennial panel survey was the 1980 decennial Census. This survey covered individuals who had at least four years of postsecondary education as of Census day 1980.

The Survey of Recent Graduates in Natural and Social Sciences and Engineering was the second major survey. In the 1980s this was a biennial cross-sectional survey of individuals who received a bachelor's or master's degree in S&E from a U.S. educational institution subsequent to Census day 1980.

The Survey of Doctorate Recipients was the final of the three surveys. This is a biennial panel survey selected from a roster of individuals who received a Ph.D. in S&E from a U.S. institution during the 42 years preceding the survey date.

The final component of the STPDS of the 1980s was a Science and Engineering Tabulation Model (SETAB). This is a computer model used to generate estimates of the total population of interest, based primarily on information from the first two surveys.

Evaluation of the 1980s STPDS

NSF recognized that there were a number of problems with the 1980s surveys, which made them less useful than desirable. For example, a last-minute funding cut led to a very hastily implemented sample size reduction of the 1982 baseline survey of experienced scientists and

engineers. In retrospect NSF realized the resulting sample design was far from optimal. Cost cutting had also led to a series of decisions which resulted in response rates for all the STPDS surveys which were considerably lower than those currently considered acceptable by the federal statistical community.

In order to ensure that similar survey design problems did not reoccur in the 1990s, NSF commissioned the Committee on National Statistics (CNSTAT) to evaluate the 1980s STPDS and make recommendations for ways in which the 1990s surveys could be improved. The CNSTAT panel consisted of experts in statistics and research design, as well as researchers who had used the data. The report from this evaluation was published in May 1989 and has been the cornerstone of NSF's redesign efforts since then.

Major CNSTAT Recommendations for the 1990s

As the title of this panel suggests, the recommendations made by CNSTAT were not for minor and gradual improvements to the STPDS. Indeed, they made recommendations for radical changes to just about every aspect of the surveys. More specifically, they recommended that NSF:

- (1) Simplify and clarify the definition of scientist and engineer.
- (2) Improve the population coverage of the surveys.
- (3) Enrich the content of the surveys.
- (4) Maximize comparability among STPDS surveys and between the STPDS and surveys conducted by other agencies.
- (5) Perform high quality state-of-the-art surveys.
- (6) Improve the sample designs of the surveys, so that sampling is more efficient.

- (7) Increase dissemination of survey results and microdata.

Current Status of Major CNSTAT Recommendations

As urged by CNSTAT, we will be clarifying our definition of scientists and engineers. In the 1980s NSF used an algorithm to define individuals' "field". Individuals were assigned a "field" based on their education, occupation, and professional self-identity. Generally, a "two out of three" rule was used, i.e., if someone had a degree in field X and was employed in field X, then (s)he was considered to be in field X. If the person's educational and occupational credentials were not consistent, field was determined by professional self-identification. Individuals were considered to be in-scope if they were in one of the physical science, social science or engineering fields of interest to NSF.

NSF's basic definition of field was not consistently defined. Rather, it was modified to take into account definitions of fields promulgated by the various professional societies. For example, to be a statistician one needed to have a Ph.D. in statistics. While the goal of this definition was to accommodate the expressed needs of STPDS users, it resulted in a complicated definition that was not consistent with definitions used by other major federal surveys.

In the 1990s NSF will present separate data series for individuals with S&E degrees and those with S&E employment. Not only will this provide users with conceptually clearer definitions, but it will allow comparability with other federal data series. It should also facilitate analyses of such issues as how do individuals with training in S&E use that training in their subsequent employment?

The CNSTAT panel also recommended that we increase the coverage of the STPDS surveys. During the 1980s our surveys coverage consisted of the following groups: (1) individuals who had at least four years of post-secondary education as of Census day and remained in the country until the reference date of the survey. (2) individuals who received a bachelor's or master's degree in an

S&E field within the two years prior to the date of the Survey of Recent Graduates. (3) individuals with a Ph.D. in S&E from a U.S. university either prior to or subsequent to Census day.

The 1980s surveys did not include individuals who received a non-S&E degree subsequent to Census day and then became employed in S&E. They also did not cover individuals who immigrated to this country after Census day unless they continued their education in this country.

In 1982 the coverage problem was slight, since the only missing groups were scientists and engineers who immigrated to this country between 1980 and 1982 and individuals who got a bachelor's or master's degree in a non-S&E field between 1980 and 1982 and then entered employment in an S&E field. By 1989 when the last attempt at estimating the total population was made, however, the size of the S&E population not included in the surveys had grown substantially. Surveys included no information for scientists or engineers who immigrated to this country between 1980 and 1989 unless they received an S&E degree subsequent to coming to this country. Likewise, we had no recent information for people who received their first bachelor's degree between 1980 and 1985. For individuals getting their first degree in 1986 or 1987 we only had survey information for individuals with a degree in science or engineering. These coverage problems contributed to our decision not to publish our 1989 estimates of the total S&E population.

While we will not be able to provide complete coverage in the 1990s, we will be able to expand coverage from the 1980s. We plan, of course, to continue the 1980s surveys - the Postcensal and panel surveys based on the Postcensal survey, the SDR, and the New Entrants surveys.

Coverage will be improved considerably by making the New Entrants survey a panel survey. This will permit us to "refresh" our experienced S&E panel with recent recipients of S&E degrees. However, it will not permit us to observe individuals who "flow" into S&E occupations without an S&E degree after Census day 1990. We will need to use information from NCES and

the Postcensal survey to estimate the size of this population.

NSF had hoped also to expand its coverage of the immigrant population. With the cooperation of INS, Oak Ridge Associated Universities conducted a pilot survey of recent immigrants for us. Based on the results of their evaluation, NSF has concluded that it is not currently feasible to design a cost-effective survey to provide information about immigrants who are educated as scientists and engineers.

The primary obstacle to the successful completion of a survey of immigrants with S&E educations is that INS does not collect information on the level and field of education of immigrants as part of their standard administrative record keeping. Therefore, to interview all those educated in S&E, we need to select cases from the entire recent immigrant population.

Our pilot survey indicated we could probably obtain a reasonable response rate from a mail/telephone survey of individuals whose INS records indicated that they had S&E or administrative occupations. However, the pilot survey response rate for the remainder of the sample was extremely low (under ten percent). While we believe we could substantially improve the response rate of the pilot survey by appropriate methodological changes, we do not believe we could obtain an acceptable response rate unless we conducted in-person interviews, using multilingual interviewers. We believe the cost of such interviews would be prohibitive.

We have initiated discussions with INS about the feasibility of their collecting information on the level and field of education of new immigrants. This would provide basic information of use to us and would increase the likelihood that we could design a cost-effective survey to obtain additional information. Such a change would also provide INS with information of relevance to the implementation of recent changes in the immigration laws.

In addition to our work on a potential survey of recent immigrants, we have been evaluating ways of making the other STPDS surveys a more fruitful source of information about immigrants and emigrants. For example, in our sample

design work we have explicitly incorporated place of birth into our stratification scheme for the Postcensal and SDR samples in order to ensure sufficient cases for analyzing this population. Until we can institute a survey of immigrants we will use estimates based on INS administrative records and information from our other surveys to estimate the size and characteristics of this population.

In summary, our population coverage will improve significantly from the 1980s. However, we will still be missing small segments of the population of interest. These segments will be larger for the population of individuals employed as scientists and engineers than it will be for the population with degrees in science and engineering. Coverage should be excellent for the population defined as having received a degree in science and engineering from a United States college or university.

The third major CNSTAT recommendation was for an enriched content in the 1990s surveys compared to the 1980s. In order to increase the usefulness of the STPDS surveys MPR and NSF staff have: (a) reviewed past STPDS questionnaires and relevant questionnaires prepared by other agencies; (b) consulted with appropriate users, STPDS contractors, and outside experts; and (c) searched appropriate literature.

We expect the 1990s instruments to provide a much richer source of information than the 1980s instruments did. However, it is never possible to do everything one would like and we have had to make some hard decisions on item selection for the surveys. The paper by Tsapogas and Gannon will provide more information on our work in this area.

The fourth major recommendation by CNSTAT was to increase comparability not only among the STPDS surveys, but also between the STPDS and other major federal surveys. In order to accomplish this goal, we have made some organizational changes. Project Officers for the STPDS surveys have been working as a group with assistance from MPR to develop survey instruments for the 1990s. This should ensure that NSF maintains a systems perspective on the surveys.

MPR and NSF staff also selected questions from other federal surveys where feasible in light of NSF's needs. When measures used in other surveys do not precisely meet our needs, we have tried to retain an ability to provide a cross-walk between our data and those of other agencies.

The greatest challenge in maintaining comparability between the STPDS data and that of other agencies has been in measuring occupation. Occupation is obviously a critical element in a survey which uses occupation to define its population of interest. However, it is also a variable that is not easy to measure.

CNSTAT recommended using a definition as close as possible to the Standard Occupation Code (SOC). However, NSF is not comfortable with simply adopting the procedures of the CPS or the decennial Census in order to define occupation. These major surveys do not always have the degree of detail desired for NSF's purposes in classifying occupation for scientists and engineers.

We do not believe that we can use the techniques of the decennial and the CPS, using finer classifications. Some of the problem is that the questions used on the CPS and the decennial do not elicit sufficient information for making the classifications we desire. For example, the overwhelming majority of postsecondary teachers in the decennial are classified into postsecondary teacher -- field not specified. For most purposes this may be adequate. NSF, however, cares a great deal whether a postsecondary teacher is teaching physics or art history.

NSF has traditionally used a closed-ended question for collecting occupational information. This forces respondents to make the distinctions in which we are interested. While we have reformulated our occupational list, so that mapping to the SOC is possible, we are very concerned that the closed list approach may lead to quite different classifications from the open-ended approaches of the CPS and the decennial. We are also concerned that the closed-ended approach will not be translatable into a CATI instrument. Because of these concerns, we have included some work on assessing occupational coding into our pretest work with the Census.

Comparability issues have extended to a number of factors other than the survey instrument. Fielding dates of the three surveys, for example, will need to be the same so that reference dates and recall periods will be the same. This has significant work load implications for us, since we previously staggered fielding dates in order to smooth out the peaks and valleys of our work.

To the extent feasible, given the differing available sampling frames for our surveys, we will use similar sample designs for the surveys in order to facilitate the combining of estimates in the SETAB model. Finally, we realize that many survey operations have potential impacts on the final estimates. For example, varying edit instructions and imputation techniques for the surveys could lead to disparate results. We are, therefore, working with MPR to provide survey contractors with standard guidelines covering these aspects of the work.

The fifth major CNSTAT recommendation was to use state-of-the-art survey techniques. Needless to say, this is a very broad and very challenging recommendation. MPR, our technical working group on methodological issues, and our survey contractors have all been assisting us in strengthening our survey techniques. The Wilkinson and Hines paper will provide more details on what we are doing in the data collection area.

Sixth, CNSTAT recommended that we use sample designs that are more efficient than those used in the 1980s. NSF staff have been working with MPR and survey contractors to develop sample designs which will balance our need for efficient estimates of the S&E population against our need to provide information on subgroups within the population and our need to minimize burden on the non-S&E population. Final designs are being developed, using computer algorithms, so that sample sizes can be cut quickly without sacrificing efficient designs in the event that budget cuts require last minute sample size changes. The Hardy paper will discuss our sample redesign work to date.

The final major area in need of improvement according to CNSTAT is increased data dissemination. We have not done a great deal of

work on meeting this goal to date, since dissemination cannot take place until data is ready. However, we have made a couple of steps in this direction.

We have reviewed our policies with respect to making the Survey of Doctorate Recipients microdata file available to researchers. While we believe that confidentiality concerns preclude broad access to the file, we have instituted procedures which can be used by researchers willing to use the data on-site at NSF. We have had two researchers take advantage of this opportunity to date and have a third researcher scheduled to use the data in the near future. In our current search for an ASA fellow, we have also included the possibility of using the soon-to-be-available 1991 SDR as an appropriate activity.

We have also started a new series of occasional papers of interest to analysts who use SRS's data to investigate specific topics. The first paper in this series, *Using the Survey of Doctorate Recipients to Measure the Number of Academic Research Personnel*, explores the impact of methodological procedures in the SDR on measures of academic research personnel.

As our work on survey design nears completion, MPR and NSF staff will turn more of their attention to data analysis and data dissemination.

The process of redesigning the STPDS surveys has proved to be a challenging one. The team working on it has frequently had to balance a great number of competing demands in order to come up with solutions to some of the problems that plagued the 1980s surveys. The work is not yet complete, but we believe that we have made a great deal of progress to date. This progress is reflected in the remaining papers in this panel.